Introduction to part 5: Corpus pragmatics

Jucker, Andreas H

DOI: https://doi.org/10.1515/9783110424928-018

18. Introduction to part 5: Corpus pragmatics

Andreas H. Jucker

1. Introduction

Part 5 of this handbook is devoted to methods in pragmatics that rely on corpus searches. Corpus pragmatics is a relatively late addition to the various subfields of pragmatics. Early work in pragmatics tended to be qualitative rather than quantitative. It tended to focus on richly contextualised instances of language use, on small sets of data and on the minutiae of spoken interaction, which precluded the use of large-scale corpora. Early work in corpus linguistics, on the other hand, tended to explore research questions in the area of lexico-grammatical, morphological and syntactic patterns and other areas of the interaction between the lexicon and sentence structure, which were amenable to be turned into search algorithms because they concerned the surface manifestations of language.

Some work in corpus pragmatics, however, appeared as early as the late 1980s and the 1990s (e. g. Aijmer 1987, 1996; Stenström and Andersen 1996; Schmied 1998 or Culpeper and Kytö 1999), but the field really took off only in the 2000s with a series of monographs and edited volumes (e. g. Aijmer 2002; Deutschmann 2003; Aijmer and Stenström 2004; Baker 2006; Facchinetti and Rissanen 2006; Adolphs 2008; Romero-Trillo 2008; Jucker, Schreier and Hundt 2009). In the meantime, the field has already matured to such an extent that in addition to a dedicated journal (*Corpus Pragmatics*) and handbook (Aijmer and Rühlemann 2015) a series of survey articles have appeared (e. g. Andersen 2011; Rühlemann 2011; Jucker 2013; Jucker and Taavitsainen 2014). Work in corpus pragmatics is proliferating at an increased pace at the moment. It combines the persisting interest in the field of pragmatics in general with the increased reliance on empirical and above all quantitative approaches and the explosion of available corpora and corpus tools (Felder, Müller and Vogel 2012; Taavitsainen and Jucker 2014).

Corpus pragmatic approaches typically adopt a quantitative perspective. Research questions often ask about the frequencies of certain elements in specific text samples and, crucially, about differences of these frequencies in different text samples. But – as I will argue in this introduction and as will become clear in the contributions assembled in this section – a quantitative perspective requires a very solid foundation in the preparation of the data base and in the analysis and categorisation of the data.

https://doi.org/10.1515/9783110424928-018
In: A. H. Jucker, K. P. Schneider and W. Bublitz (eds.). *Methods in Pragmatics*, 455–466. Berlin/Boston: De Gruyter Mouton.

2. The scope of corpora

In a pre-theoretical sense, any collection of texts or even one single text can be called a corpus. In the sense intended here, however, only electronically searchable corpora are meant. In the definition of Andersen (2011: 590), "corpora are compilations of naturally occurring spoken or written language that can be accessed on a computer. Such compilations may be monolingual or multilingual and may represent general language or specific domains (professional/academic corpora)".

The earliest corpora in this sense date back to the 1960s. They were designed to provide a more or less representative mirror image of an entire language, and a lot of thought went into the balanced construction of these corpora: which text genres should be represented? And how should the different genres be distributed? According to Aarts' (2011) useful typology, such corpora are, therefore, called balanced corpora. Examples of such early balanced corpora are the *London-Lund Corpus of Spoken English* (LLC), the *Brown Corpus* of written American English or the *Lancaster-Oslo-Bergen* (LOB) Corpus of written British English. Aarts (2011) stresses the intuitive nature on which the "balancing" was done. There is, as yet, no established way to assess in any useful sense the overall composition of a language as a whole, and, therefore, it can only be pure guess work what kind of composition of a sample corpus would best represent an entire language. To a large extent this is also true for specialised corpora that try to represent a single variety of a language. The corpus of *Early Modern English Medical Texts* (EMEMT), for instance, claims to be a "representative sample of the entire field of English medical writings that appeared in print between 1500 and 1700" (Taavitsainen and Pahta 2010: cover blurb). However, from a strictly statistical point of view, such a claim rests on a full and comprehensive list of all the relevant texts of the entire field and a selection principle which gives every single text of the field the same chance of being included in the sample corpus, a criterion which seems hard to achieve even in a limited field such as medical discourse. In the case of an entire language, there is no way of establishing the limits of the entire set (or "population" in statistical terms) that a corpus is supposed to represent. Corpora still try to be representative of more than just themselves, and, therefore, the label "sample corpus" seems more appropriate according to Aarts (2011). He mentions the *British National Corpus* with 100 million words as the largest sample corpus of British English.

According to Aart's (2011) typology, there are also full-text corpora, which contain one or more complete texts. Parallel corpora contain texts of more than one language or more than one variety of the same language. The parallelism between these texts can vary from direct translations of one language into the other to corpora of different varieties or languages that have been compiled on the basis of identical designs. The Brown and LOB corpora, for instance, consist of identical samples of different genres drawn from American English and British English respectively. Additional categories are diachronic or historical corpora represent-

ing older stages of a language and learner corpora containing texts produced by non-native speakers of a language.

In recent years, the number of available corpora and their size have increased at an unprecedented rate. Back in the 1960s one-million-word corpora were considered to be large. In the meantime, many corpora are available extending to several hundred million words. A dedicated website created by Mark Davies includes a dozen different corpora, four of which contain more than one billion words (http://corpus.byu.edu). It includes balanced corpora such as the *Corpus of Contemporary American English* (COCA, 520 million words) but also corpora with a very narrow focus on just one type of text, e. g. the *Hansard Corpus* with the proceedings of the British Parliament from 1803 to 2005 (1.6 billion words) or the *Corpus of American Soap Operas* with transcripts from American soap operas from the early 2000s (100 million words). The largest corpus, however, is provided by the *Google Books Ngram Viewer*, which accesses a database of 361 billion words.

However, for research questions in pragmatics, corpus size is usually not the decisive criterion. It is usually more important for the pragmaticists to be able to contextualize the individual search results, either in the immediate context surrounding the search item or the larger context of the genre or text type in which it occurs. The Ngram Viewer does not provide any context at all. In fact, the searches are not performed on entire texts but on indexes derived from the texts. The ngrams in these indexes carry only minimal information about the type of English and the year of publication of the text in which they originally occurred. In other corpora, it is usually possible to trace individual occurrences of search items back to their original location but often this has to be done manually, which severely restricts the amount of data that can be assessed in this way in spite of the ease of retrieving many more occurrences from these large corpora. Thus, there is often a tension between small but richly contextualised sets of data versus large-scale corpora with a lot of quantifiable material but a very limited amount of context for each of the retrieved hits; the big data caveat in O'Keeffe's terms (this volume; see also Taavitsainen and Jucker 2015: 18).

One solution to this problem is the use of pragmatically annotated data (see Archer and Culpeper, this volume). A subcorpus of the *Michigan Corpus of Academic Spoken English* (MICASE), for instance, has been tagged for some speech acts, and the *Corpus of Verbal Response Mode* (VRM) *Annotated Utterances* has been coded both for literal meaning and for pragmatic meaning (see Rühlemann 2011: 630). But such annotations are extremely labour intensive, which puts severe limitations on the size of the corpora that can be annotated in this way.

3. Corpora, quantification and statistics

Corpus pragmatic approaches search for patterns and generalisations across large amounts of data. Research questions typically ask for frequencies and differences in frequencies in different samples or subsamples. They ask questions that can only be answered with numerical results. However, any numerical claim depends on a solid foundation consisting of several layers pertaining to the database, the identification and analysis of the data and so on. This can be visualised as a pyramid in which each individual level depends on a solid foundation of all the lower levels, and at the same time each level consists of a higher degree of abstraction and generalisation than its supporting level and thus the height of each level comes at the cost of a further loss of detail (see Figure 1).

Figure 1 depicts the pyramid of quantitative research. At the bottom of any quantitative research there is the selection and compilation of data. The researcher can decide to make use of an existing corpus or to construct a corpus specifically designed for the research question at hand (see chapter 19 by Gisle Andersen). The decision is not trivial. Mistakes at this level may render all the work at higher levels questionable or even meaningless. Considerations at this level will include the question about which language varieties need to be included, whether they are spoken or written, the degree of formality, the diachrony of the data and many more. The second level of the pyramid very often consists of the pre-processing of the data (see chapter 20 by Dawn Archer and Jonathan Culpeper). Present-day corpora are often annotated with parts-of-speech tags. There are also speaker-identification tags and tags that identify different registers or modalities of the language samples that are included. Some corpora even include pragmatic annotations. The quality of these annotations again has an immediate bearing on the reliability of all the work carried out at the higher levels in the pyramid. If the accuracy of the parts-of-speech tags is less than one hundred per cent, for instance, the quantifications at the higher levels inherit these errors to the extent that they rely on the parts-of-speech tagging.

The core of any research project is, of course, the identification and description of a certain linguistic phenomenon. In the context of corpus pragmatic research this can be a particular linguistic form or a range of such forms, such as a particular discourse marker or an interjection, whose functions are to be investigated (see chapter 21 by Karin Aijmer), or a range of speech functions, such as a specific speech act or a class of speech acts, whose specific linguistic realisations are to be investigated (see chapter 22 by Anne O'Keeffe). A precise description of these phenomena is again an indispensable prerequisite in order to ensure the reliability of the higher levels in the pyramid.

Once the elements have been identified, they need to be categorised. Different uses of a discourse marker, for instance, or specific ways of realising a certain speech act have to be distinguished. Without such a categorisation, the elements

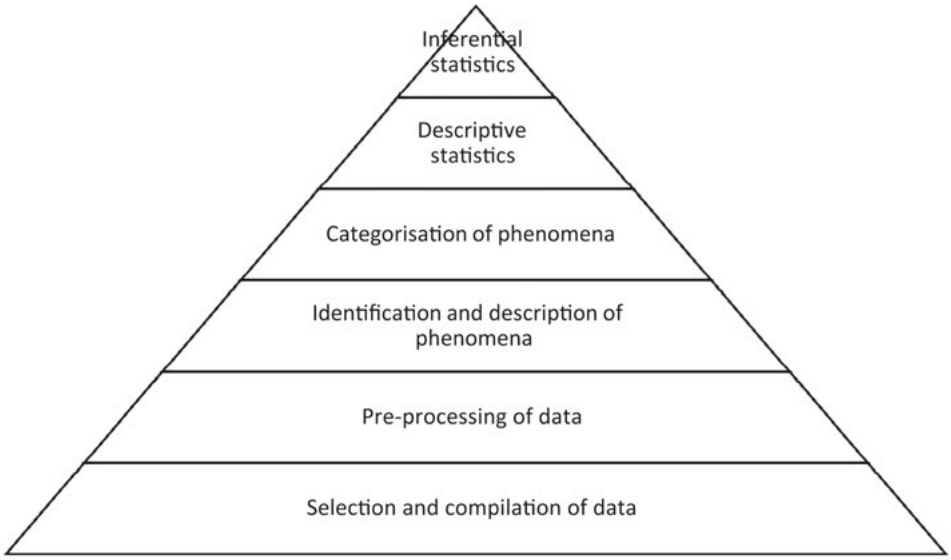

Figure 1: The pyramid of quantitative corpus research

cannot be counted and quantification is not possible. The items to be quantified need to be identified in such a way that they can be reliably counted. This means that individual occurrences of some phenomenon are claimed to be sufficiently similar or even identical in order to be lumped together. Small differences that are not relevant are abstracted away or ignored. In this sense quantification necessarily involves a certain loss of detail of description. It is the price that has to be paid for quantification. If we are prepared to pay the price, we can count the instances, and we can compare different phenomena.

It is also essential at this stage – and this is all too often ignored – that the categories must be defined in such a way that another researcher would identify the same elements as instantiations of this particular category. This stage, therefore, should include an interrater reliability test. This involves at least two raters, or coders, who independently code a data sample and then compare their results. The categorisation is only considered to be sufficiently robust if the coders come up with a sufficiently high number of identical codes assigned to the data. If that level is not achieved, the category descriptions have to be improved or the categories have to be adjusted before a new round of testing with fresh data samples can be started. This process has to be repeated until the desired level of interrater agreement has been achieved. Usually a level of 70 per cent is considered to be adequate. Practical experience shows that such a level, which may appear to be relatively modest, is often more difficult to achieve than might be expected, especially if functional categories are involved. However, the reliability of category

counts critically depends on the reliability of category identification. If the categories proposed by the researcher have not passed the test of interrater reliability, the quantitative results have to be seen with a lot of scepticism, and even if they have passed such a test, it should be clear that a level of a minimum of 70 per cent interrater agreement means that the results are no more than approximations or relatively accurate estimates. The nature of linguistic data generally does not lend itself to high precision measurements.

This scepticism is essential whenever higher levels in the pyramid are considered. The counting of categories that forms the basis for the descriptive statistics seems like a tedious task that can generally be done easily and quickly by the computer. But the ease of computation should not be allowed to suggest a degree of precision that is not supported by the approximate nature of the underlying data categorisation.

At the level of descriptive statistics, researchers often have to work with normalised frequencies. If the frequencies of a certain linguistic element are to be compared in two or more different contexts, the actual figures have to be set into relation of the size of these contexts. Normally this is done in terms of number of words. The observed frequency of the element in each context is calculated as a frequency per 10,000 words or per one million words or some other suitable level. It seems straightforward to use the number of words as the category for normalisation but it is not without problems. Computers can count the number of words very easily and quickly but they rely on a rather crude definition of what a word is (something like a string of letters enclosed by blanks or punctuation marks). Even if this is too simplistic for a linguistic definition of what a word is, for many purposes it is good enough as a proxy, in particular if the word count is carried out in the same way in all the relevant contexts. But in some instances the number of turns or the duration of speaking may be more accurate measures for the normalisation of frequency figures, and it must be realised that the results depend on such choices.

The pinnacle of many research efforts seems to be reached when the researcher cannot only produce the frequencies for a particular element in different contexts but when he or she can confidently claim that the differences are significant. This is done on the basis of inferential statistics. Many different statistical tests are available for this purpose, and the computer will very quickly return a verdict of whether different numerical patterns in the different contexts are likely to be random or whether they are sufficiently large to exclude the possibility of being just random and, therefore, must be assumed to be significant.

However, such results must always be addressed with a healthy dose of caution. It crucially depends on the choice of an appropriate statistical test, and it depends just as crucially on the reliability of the figures that have been fed into the computer, which depends – as argued above – on the quality of the choices at all the lower levels of the pyramid. But even with the best of intentions and the highest level of care, the result at the top of the pyramid inherits all the unavoidable

limitations at the lower levels. It only applies to the data that was included in the sample, it depends on the accuracy of the data annotations, the reliability of the data categorisation and counting, and so on.

And ultimately, even if we accept – with sufficient caution – the significance of our results, the statistical tests do not tell us anything about the reasons for this significance. A distribution of the data that is highly unlikely to be random is just that – a distribution that is highly unlikely to be random – no more, no less. Often enough it is just the starting point for new questions to be asked.

4. The papers in this section

The first two papers in this section are concerned with the construction and annotation of corpora. In chapter 19, Gisle Andersen discusses the various aspects that need to be taken into consideration when researchers either choose an existing corpus or decide to build their own corpus. He argues that the specifics of pragmatic research often make it useful or even indispensable to go beyond ready-made, off the shelf corpora by either extracting relevant subparts, by annotating existing corpora in various ways or by embarking on the construction of the researcher's own tailor-made corpora. Andersen focuses on the various selective processes, or sampling frames, of corpus construction and on the effects these choices have on the potential for corpus pragmatic investigations. He discusses the differences between form-based approaches and function-based approaches and the distinction between corpus-based versus corpus-driven approaches. The sampling frame is particularly challenging in the case of parallel corpora with data drawn from different languages or different time periods because the inventory of genres and text types may be very different in these languages or time periods. He also discusses some more technical aspects of corpus construction, such as the transcription of spoken data and various types of annotations.

In chapter 20, Dawn Archer and Jonathan Culpeper argue that pragmatic annotation for a long time lagged behind the annotation of other aspects in corpora. They note that corpus pragmatic work so far has had a strong bias towards research questions with a formal entity as a starting point. Pragmatic annotation offers a way out of this restriction. They distinguish between different levels of pragmatic annotation. At one level, there are annotation schemes that identify interactional phenomena, such as speech acts, and at a second level, there are annotation schemes for contextual phenomena, such as the gender or social status of the interactants. Such contextual features are particularly important since pragmatic interpretations are regularly based on contextual features. The annotation of pragmatic units is difficult because of the problem of identifying adequate boundaries and because pragmatic units are often ambiguous and indeterminate. Pragmatic annotations, therefore, must often be applied manually, which seriously restricts the corpus

size for annotations. They also present their own annotation scheme, which they used for the *Sociopragmatic Corpus* with its sophisticated and highly detailed tags identifying for each segment the relevant combination of sociopragmatic variables including speaker identification, addressee identification, and their relationship. They argue that many pragmatic phenomena cannot easily be annotated automatically but some annotation is possible with computational assistance.

The third paper in this section, by Irma Taavitsainen, chapter 21, is devoted to the historical dimension of corpus pragmatics, where the challenges and problems of corpus pragmatic research are exacerbated because of the historical nature of the data. She provides an outline of the relevant corpora, from the pioneering *Helsinki Corpus* to the single-register or single-variety corpora produced by the same Helsinki team to more recent corpora. She focuses on some of the challenges of historical corpus pragmatic work, such as the dilemma between large generalisations which cover a lot of data versus the wish to focus on increasingly fine-grained distinctions, which reduces the available data for each relevant distinction to such an extent that useful generalisations are no longer possible, or the problem of spelling variation in historical texts. The chapter also gives a brief introduction to the most important corpus tools, such as concordances, keyword analysis, collocations and statistical assessments, and it points out the importance of including the social and cultural context as well as the genre context into the analysis. This makes it necessary to switch back and forth between the frequency counts of corpus searches and the actual contexts in which the search items occur. Finally, she identifies some future directions for historical corpus pragmatics, as for instance an increased trend towards megacorpora, towards increasingly richer and more sophisticated annotations of corpora, and towards more and more sophisticated editing techniques that are used to prepare historical material for inclusion into searchable corpora.

Chapters 22 and 23 consider the relationship between form and function in corpus pragmatics. The chapter by Karin Aijmer looks specifically at research approaches that take a linguistic form, such as a discourse marker, an interjection, a term of address or a hesitation marker as a starting point in order to explore its function across a large number of occurrences. This is the more common approach in corpus studies because corpus searches depend on clearly specifiable strings of linguistic material, i. e. on formal patterns. She draws attention to the problem of the ambiguity of many linguistic forms. Discourse markers, for instance, often have linguistic forms that coincide with forms in other word classes and even as discourse markers they are multifunctional. She, too, draws attention to the importance of the context for the interpretation of the various functions of the elements retrieved in corpus searches. She also points out the connection to the variationist perspective, in which search items are systematically correlated with different types of context in order to explore the sociolinguistic factors, for instance, on the usage of specific elements. Moreover, she considers corpus pragmatic work in the context of selected theoretical approaches, such as Thetical Grammar or Construction Grammar.

The paper by Anne O'Keeffe looks at approaches that take a speech function, e. g. a specific speech act, as a starting point in order to explore its realisations in a specific set of texts. This can be done by searching for elements that are regularly associated with this function, as for instance *sorry*, which may function as an apology or may accompany an apology. But not all apologies contain an instance of *sorry*, and not all instances of *sorry* occur together with an apology. She also draws attention to the dilemma in corpus research between large numbers of occurrences of a particular phenomenon, breadth of forms in her words, and the contextual depth that is available for each occurrence. The larger the number of occurrences, the more restricted will be the contextual depth for each occurrence and vice versa. In order to illustrate the problem, she traces the history of *I'm sorry* and *I apologise* in the largest available corpus, the *Google Books Ngram Viewer*. She then presents two case studies which contrast corpus linguistic methods and discourse completion tasks. The study by Schauer and Adolphs (2006), which analyses expressions of gratitude in the *Cambridge and Nottingham Corpus of Discourse in English* (CANCODE) and in a discourse completion task, finds that the corpus data gives a broader contextual picture than the DCT data. In the corpus, expressions of gratitude often occur in clusters while in the DCT data single utterances expressing gratitude are the norm. This result is supported by a study by Flöck and Geluykens (2015), who compared directives in the British component of the *International Corpus of English* (ICE) with response data of a written DCT and a small corpus of business letters. In the final part of the chapter, O'Keeffe presents different approaches that deal with the problem of searching for speech functions. The first approach, one-to-one searching, is restricted to instances in which a specific form, such as *thank you*, or a specific tag is searched for. This will provide a full recall of all such forms. The second approach consists of a down-sampling of the corpus to a manageable size and a manual analysis of the relevant search item. The third approach makes systematic use of existing research findings, e. g. from DCT studies, to establish the relevant search items for corpus search. And finally she presents four possible solutions that have been proposed for larger corpora together with their advantages and limitations: the use of illocutionary force indicating devices; the use of genre-specific search inventories established by manual searches of small sample corpora; the use of typical lexical or grammatical features associated with a speech act; and, finally, the use of metacommunicative expressions.

In the last chapter of this volume, chapter 24, finally, Michael Haugh focuses specifically on the corpus-pragmatic approaches that take metapragmatic elements as a starting point. Such elements reflect the interactants' awareness of what is going on in the interaction and their comments about this. Haugh uses elements such as *just kidding*, *kidding*, *only joking* and so on as examples with which the speaker signals to the addressee that the surrounding talk should be treated as non-serious, playful or jocular. He distinguishes between three different types of acts and activities: first, pragmatic acts and activities (e. g. *apologise, joke,*

threaten); second, inferential acts and activities (e. g. *allude*, *imply*, *sarcasm*); and third, evaluative acts and activities (e. g. *aggressive*, *polite*, *rude*). He identifies a number of challenges of an analysis of metapragmatic elements. First, the analysis must identify a sufficient number of tokens for an analysis, and these tokens must be comparable across contexts. The same metapragmatic lexical item may well be used in different ways on different occasions. And second, the accuracy of the transcriptions is essential. A careful transcription often reveals details that are lost in a less detailed rendering.

Part 4 of this handbook covered methods that were largely qualitative. They focused on small data sets of richly contextualised communicative behaviour. In the following chapters of part 5 of the handbook, the focus shifts to large scale investigations that try to find generalisations across ever increasing data sets. But the tension between such large-scale generalisation and the goal of paying attention to the minute details of each individual occurrence remains a *leitmotif* in all the chapters of part 5.

References

Aarts, Jan
2011 Corpus analysis. In: Jan-Ola Östman and Jef Verschueren (eds.), *Handbook of Pragmatics Manual*. Amsterdam/Philadelphia: John Benjamins.

Adolphs, Svenja
2008 *Corpus and Context. Investigating Pragmatic Functions in Spoken Discourse*. (Studies in Corpus Linguistics 30.) Amsterdam: John Benjamins.

Aijmer, Karin
1987 *Oh* and *ah* in English conversation. In: Willem Meijs (ed.), *Corpus Linguistics and Beyond. Proceedings of the Seventh International Conference on English Language Research on Computerized Corpora*, 61–86. Amsterdam: Rodopi.

Aijmer, Karin
1996 *Conversational Routines in English. Convention and Creativity*. London: Longman.

Aijmer, Karin
2002 *English Discourse Particles. Evidence from a Corpus*. (Studies in Corpus Linguistics 10.) Amsterdam: John Benjamins.

Aijmer, Karin and Christoph Rühlemann (eds.)
2015 *Corpus Pragmatics. A Handbook*. Cambridge: Cambridge University Press.

Aijmer, Karin and Anna-Brita Stenström (eds.)
2004 *Discourse Patterns in Spoken and Written Corpora*. (Pragmatics & Beyond New Series 120.) Amsterdam: John Benjamins.

Andersen, Gisle
2011 Corpus-based pragmatics I: Qualitative studies. In: Wolfram Bublitz and Neal R. Norrick (eds.), *Foundations of Pragmatics*, 587–627. (Handbooks of Pragmatics 1.) Berlin: de Gruyter Mouton.

Baker, Paul
2006 *Using Corpora in Discourse Analysis*. London: Continuum.
Culpeper, Jonathan and Merja Kytö
1999 Modifying pragmatic force: Hedges in a corpus of Early Modern English dialogues. In: Andreas H. Jucker, Gerd Fritz and Franz Lebsanft (eds.), *Historical Dialogue Analysis*, 293–312. Amsterdam: John Benjamins.
Deutschmann, Mats
2003 *Apologising in British English*. (Skrifter från moderna språk 10). Umeå: Institutionen för moderna språk, Umeå University.
Facchinetti, Roberta and Matti Rissanen (eds.)
2006 *Corpus-based Studies of Diachronic English*. (Linguistic Insights 31.) Bern: Peter Lang.
Felder, Ekkehard, Marcus Müller und Friedemann Vogel (Hrsg.)
2012 *Korpuspragmatik. Thematische Korpora als Basis diskurslinguistischer Analysen*. Berlin: de Gruyter.
Flöck, Ilka and Ronald Geluykens
2015 Speech Acts in corpus pragmatics: A quantitative contrastive study of directives in spontaneous and elicited discourse. In: Jésus Romero-Trillo (ed.), *Yearbook of Corpus Linguistics and Pragmatics 2015*, 7–37. London: Springer.
Jucker, Andreas H.
2013 Corpus pragmatics. In: Jan-Ola Östman and Jef Verschueren (eds.), *Handbook of Pragmatics*, 2–17. Amsterdam: Benjamins.
Jucker, Andreas H., Daniel Schreier and Marianne Hundt
2009 Corpus linguistics, pragmatics and discourse. In: Andreas H. Jucker, Daniel Schreier and Marianne Hundt (eds.), *Corpora: Pragmatics and Discourse. Papers from the 29th International Conference on English Language Research on Computerized Corpora (ICAME 29). Ascona, Switzerland, 14–18 May 2008*, 3–8. (Language and Computers: Studies in Practical Linguistics 68.) Amsterdam: Rodopi.
Jucker, Andreas H. and Irma Taavitsainen
2014 Diachronic corpus pragmatics: Intersections and interactions. In: Irma Taavitsainen, Andreas H. Jucker and Jukka Tuominen (eds.), *Diachronic Corpus Pragmatics*, 3–26. (Pragmatics & Beyond New Series 243.) Amsterdam: John Benjamins.
Romero-Trillo, Jesús (ed.)
2008 *Pragmatics and Corpus Linguistics. A Mutualistic Entente*. (Mouton Series in Pragmatics 2.) Berlin: Mouton de Gruyter.
Rühlemann, Christoph
2011 Corpus-based pragmatics II: Quantitative studies. In: Wolfram Bublitz and Neal R. Norrick (eds.), *Foundations of Pragmatics*, 629–656. (Handbooks of Pragmatics 1.) Berlin: de Gruyter Mouton.
Schauer, Gila A. and Svenja Adolphs
2006 Expressions of gratitude in corpus and DCT data: Vocabulary, formulaic sequences, and pedagogy. *System* 34: 119–134.
Schmied, Josef
1998 Discourse markers in the Lampeter Corpus of Early Modern English Tracts. In: Raimund Borgmeier, Herbert Grabes and Andreas H. Jucker (eds.), *Anglistentag 1997 Giessen. Proceedings*, 57–65. Trier: Wissenschaftlicher Verlag.

Stenström, Anna-Brita and Gisle Andersen
1996 More trends in teenage talk: A corpus-based investigation of the discourse items *cos* and *innit*. In: Carol E. Percy, Charles F. Meyer and Ian Lancashire (eds.), *Synchronic Corpus Linguistics: Papers from the Sixteenth International Conference on English Language Research on Computerized Corpora (ICAME 16)*, 189–203. Amsterdam: Rodopi.

Taavitsainen, Irma and Andreas H. Jucker
2015 Twenty years of historical pragmatics: Origins, developments and changing thought styles. *Journal of Historical Pragmatics* 16(1): 1–25.

Taavitsainen, Irma and Päivi Pahta (eds.)
2010 *Early Modern English Medical Texts. Corpus Description and Studies*. Amsterdam: John Benjamins.